STAT 672 Statistical Learning & Data Analytics Spring 2018

George Mason University

Instructor: Martin Slawski

Homework 4

Pencil-and-paper part due Thursday, April 5th, in class. Code and relevant output has to be submitted via Blackboard. Please follow the following format for the filename of your submission: `Lastname_Firstname_HWx.zip`, where `x` needs to be substituted by the homework #.
Maximum possible points to earn: 18; 100% = 15 points, rest counts as bonus.

Soft-margin SVM

In this problem, we apply the (linear) soft-margin SVM to the `spambase` data set from the UCI machine learning repository (`https://archive.ics.uci.edu/ml/datasets/Spambase`). Please have a look at that page and follow the link to "Data Set Description" to get more background information about the data set. Then download the files `spam_Xtrain.csv`, `spam_Ytrain.csv`, `spam_Xtest.csv`, `spam_Ytest.csv` from blackboard. The first two files contain the features respectively labels of the training set, and accordingly the remaining files form the test set. The file `featurenames` contains the names of the features.

In order to work on this homework problem, you are supposed to use (an interface to) `LIBSVM` (`https://www.csie.ntu.edu.tw/~cjlin/libsvm/`).

`Python`: scikit-learn
`R`: package `e1071`.

a) Read in the data. Center and scale `Xtrain` and `Xtest` as follows:

$$X_{ij} = \frac{X_{ij} - \bar{x}_j}{\zeta_j}, \;\; i \in \text{train and } i \in \text{test},$$

$$\text{where } \bar{x}_j = \frac{1}{n_{\text{train}}} \sum_{i \in \text{train}} X_{ij}, \quad \text{and } \zeta_j = \left(\sum_{i \in \text{train}} (X_{ij} - \bar{x}_j)^2 \right)^{1/2}.$$

Why is it recommended to scale the data when using the SVM?

Hint. Use the connection between the SVM and regularized empirical risk minimization.

b) Run the linear support vector machine with input with cost parameter $C \in \{10^{-1}, 10^{-0.5}, \ldots, 10^{5.5}, 10^{6}\}$.

Please make sure that all input parameters are set correctly (kernel = “linear", no additional scaling, ...).

For each of the above values of C, record the following quantities:

- The number of support vectors,
- the training error,
- the test error (for the centered + scaled test data as detailed in a)),
- the false positive and true positive rate for the test set,
- the optimal weight vector $w^* = \sum_{i \in \mathcal{S}^*} x_i \alpha_i^* y_i$, where the $\{\alpha_i^* \geq 0\}$ are the optimal dual coefficients, and $\mathcal{S}^*$ is the set of support vectors.

c) How does the number of support vectors behave as C is increased? Explain. How does the time needed for training behave as C is increased (no time measurements required)? When increasing C further, you will observe that the solver does not terminate. Explain.

d) How can the entries of w^* be interpreted at the level of the features? For $C = 10^{4.5}$ find the variable names corresponding to the largest five positive and the largest five negative entries of w^*. Interpret what you find.

e) Inspect the false positive and true positive rates. Suppose that it is tolerable to have about 1% of actual email wrongly classified as spam – what would be the best choice of C (among those values used in b))?

f) For the problem under consideration, different forms of misclassifications (spam classified as email, email classified as spam) are naturally associated with different costs as it is highly undesirable to miss important email. Explain how the soft-margin SVM optimization could be modified to incorporate different cost for each type of misclassification.

Hint. Be explicit – argue directly in terms of the optimization problem of the soft margin SVM.

Point distribution:

a)	b)	c)	d)	e)	f)
4	5	2	3	2	2